# 3. Statistical Confidentiality Process

Statistical confidentiality is the process used to reduce the risk of data disclosure through the release of survey data. This process includes at least one of the following: (1) a set of statistical methods that reduce or modify the data and (2) restricting the accessibility of the data through various [data release](#CHAPTER_2_DATA_RELEASE_OPTIONS) options.

The process is aimed at reducing the risk that an “intruder” with potentially identifying information about a data subject can match the information to that available in the released file. The intruder could thereby identify an individual unit in the released file and disclose additional data collected in the survey. Presumably, the survey data contain information to which the intruder does not have access initially and which should not be disclosed. Possible intruders include a survey respondent’s spouse, parent, child, neighbor, or other individual.

[*In preparing and/or releasing the survey data, you must evaluate the statistical disclosure risks so you can plan and/or confirm the existence of sufficient protection for the confidential information of various data subjects (e.g., schools, students). The disclosure risk analysis requires a good understanding of potential intruders, including their resources of identification (e.g., matching keys and accessibility to databases). These aspects will influence the confidentiality strategies or methods you choose for your project.*

*The statistical confidentiality strategies are also influenced by the modes of data release. See section 2.2 for more guidance on data release.*

*Why the concern about confidentiality? The motivation for the confidentialityprocess is the set of laws that dictate the requirements for confidentiality. There are several laws of which you should be aware, and links to some of the more general laws are provided below. In addition to the laws, a motivating factor is to keep the promise of confidentiality made to respondents at the time of interview. Violation of that promise will result in reduced trust by data subjects, as well as lower response rates.*]

The general statistical confidentiality process begins with an initial risk analysis, discussed in Section 3.1. Section 3.2 continues with a description of the confidentiality treatments that are needed prior to releasing data through the planned data dissemination modes. Section 3.3 provides a discussion of the rules for reporting.

## 3.1 Initial Risk Analysis

The purpose of conducting an initial disclosure risk analysis is to inform the statistical confidentiality process for each data file. The initial risk analysis will, at the very least, identify combinations of variables that may help to identify a respondent when considered together.

* *[Run frequencies/cross-tabulations for every variable on the file. Temporarily categorize any continuous variables and run frequencies for those variables as well. Look for any categories that are especially sparse.*
* *Gradually add variables to the frequencies (one at a time) to identify any problems. Note that when frequencies involve more than three or four variables, most records will look* [*unique*](#Population_unique/sample_unique)*.*
* *Perform cross-tabulations for all of the key variables.*
* *If cross-tabulations with a particular variable continually result in sparse cells or unique records, then that variable may need to be collapsed further. However, keep in mind the* [*methods that reduce risk by modifying microdata*](#_4.3.1_Methods_That)*, which may eliminate the need for further collapsing.*
* *Look for outlying values in the continuous variables. They may need to be modified in some manner before being released on the data file (e.g., by top-coding).*
* *You might also consider a more formal risk modeling approach in which you identify key variables and then estimate level of risk for each record. Skinner and Holmes (1998), for example, suggest a risk-model approach for estimating the union uniqueness (i.e.,* [*unique*](#Population_unique/sample_unique) *in both the sample and the population) with a set of categorical key variables. This approach is implemented in the* [*µ ARGUS software*](#ARGUS) *(2003).]*